School of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies (Semester 1, 2019)
Workshop sample solutions: Week 9

1. What is **Machine Learning** (or **Data Mining**) and why is it a **Knowledge Technology**? Why has this field become important? Contrast the use of data/information/knowledge/wisdom with the way we have used them previously in this subject.

   - "Data mining" is a blanket term referring to a set of knowledge technologies problems where the analysis of quantifiable data leads to inferences about quantifiable (typically statistical) observations or patterns about the data.
   - "Machine learning" is itself a cover term for a set of algorithms (like classification, clustering, or regression algorithms) for mechanically manipulating our data set to produce certain statistical observations; usually used as part of data mining
   - These are important because of the explosion in the quantity of available data, where useful information is encoded — this has increased far more quickly than our capacity for managing data, and getting knowledge out of it (as humans)

2. An urn initially contains 10 red balls and 6 black balls. At each trial, a ball is selected from the urn, its colour is noted, and then it is returned to the urn with 2 more balls of the same colour.

   - There are two events: the first trial T_1 and the second trial T_2, each of the trials can show either a red ball R or a black ball B.

   (a) Compute the probability of obtaining a red ball with both the first and second trials

   - First, we find the probability of observing a red ball on the first trial:

$$\begin{aligned} P(T_1 = R) &= \frac{\text{\# of successful instances}}{\text{total \# of instances}} \\ &= \frac{10}{16} = \frac{5}{8} \end{aligned}$$

   - Then, we notice that, if we observe a red ball the first time around, the distribution in the urn for the second trial changes to 12 red balls and 6 black balls. So, the probability of observing a red ball on the second trial, *given* that we have observed a red ball on the first trial, is $\frac{12}{18}$. Now we can find the joint probability (using the **product rule**):

$$\begin{aligned} P(T_1 = R \cap T_2 = R) &= P(T_2 = R \mid T_1 = R) \cdot P(T_1 = R) \\ &= \frac{12}{18} \cdot \frac{10}{16} \\ &= \frac{5}{12} \end{aligned}$$

   (b) Show that the events "red ball on the first trial" and "red ball on the second trial" are **not** independent

   - If two events are independent, then the joint probability that they both occur is equal to the product of the prior probabilities of each event occurring. In this case, that would be:

$$P(T_1 = R \cap T_2 = R) = P(T_1 = R) \cdot P(T_2 = R)$$

   - The left hand side is $\frac{5}{12}$ from (a). We also found that $P(T_1 = R) = \frac{10}{16}$ in (a). As for $P(T_2 = R)$, we need to consider both of the cases: where the first trial was red, and where the first trial was black (the (normalised) **sum rule**):

$$\begin{aligned} P(T_2 = R) &= P(T_2 = R \cap T_1 = R) + P(T_2 = R \cap T_1 = B) \\ P(T_2 = R) &= P(T_2 = R \mid T_1 = R) \cdot P(T_1 = R) + P(T_2 = R \mid T_1 = B) \cdot P(T_1 = B) \\ &= \frac{12}{18} \cdot \frac{10}{16} + \frac{10}{18} \cdot \frac{6}{16} \\ &= \frac{5}{8} \end{aligned}$$

- So, the right-hand side of the independence equation above is $\frac{5}{8} \cdot \frac{5}{8} = \frac{25}{64}$, and the left hand side is $\frac{5}{12} = \frac{25}{60}$. These are clearly not equal, so the events must not be independent.

3. Revise the definition of **data mining**. Name the main modes of data mining, and give an example of a task for each one.

   - **Classification**, for example, choosing the best label which represents the topic of a text document
   - **Clustering**, for example, finding groups of network attackers which have something(s) in common
   - **Regression**, for example, predicting the numeric quantity representing the sea level rise in 2030
   - **Association Rule Mining**, for example, finding which purchases are strongly predictive of other purchases in a credit card dataset
   - **Sequential Discovery**, for example, finding weather patterns in a given city, or how a catalyst will change the rate of a chemical reaction
   - **Outlier Detection**, for example, finding one or more erroneous values in an otherwise regular collection of data